Méthodes économétriques de décomposition des inégalités

De la théorie à la pratique

Sophie Maillard (Insee)
Béatrice Boutchenik (Insee, Université Paris-Dauphine)

Journées de méthodologie statistique - 12 juin 2018

Les méthodes de décomposition

- Un ensemble de méthodes parmi les plus importantes pour l'analyse des différences entre groupes bien identifiés. Applicable aux inégalités de genre... mais pas seulement !

- Des outils mobilisables dans toute situation où l'on cherche à comprendre **l'écart entre deux groupes mesuré selon une variable d'intérêt donnée, au regard des différences de caractéristiques entre les deux groupes**.

- En passe de devenir un **outil de politique publique** - logiciel de calcul au sein des entreprises des écarts inexpliqués entre H et F et obligation de les résorber dans un délai de 3 ans (article 61 du projet de loi pour la liberté de choisir son avenir professionnel).

Évolution du questionnement dans la littérature économique

Oaxaca (1973) et Blinder (1973)

- Écart de salaires hommes/femmes et blancs/noirs aux US
- Décomposition entre salaires moyens
- Un des outils les plus fréquemment utilisés en économie du travail.

Plus récemment : attention particulière portée à la décomposition des inégalités, au-delà de la moyenne

- Évolution temporelle de la distribution des revenus
- Étude plus fine des discriminations (*glass ceiling* ou *sticky floor*)

Une approche pratique

- Référence théorique indispensable : N. Fortin, T. Lemieux, S. Firpo (2011) Decomposition Methods in Economics. *Handbook of labor economics*, 4, 1-102.

- Comment mettre en oeuvre ces méthodes ? Quelles questions se poser à chaque étape ? Comment interpréter les résultats ?

- Document de travail méthologique en préparation et une deuxième session de formation fin 2018.

Principes des méthodes de décomposition

La décomposition des écarts de moyenne

La décomposition de l'écart entre distributions

Les régressions RIF

Principes des méthodes de décomposition

Idée générale

- On compare deux groupes suivant une variable d'intérêt (dont on considère une certaine statistique). On suppose qu'ils diffèrent en deux aspects :
  - Les caractéristiques déterminant le niveau de la variable d'intérêt ne sont pas distribués de la même façon dans les deux groupes (composition).
  - Le lien entre ces caractéristiques et la variable d'intérêt (la valorisation de ces caractéristiques) n'est pas le même entre les groupes (valorisation).
- L'objectif est de mesurer les effets de ces deux mécanismes, au travers d'un écart dit **expliqué** et un écart **inexpliqué**.

A quoi correspondent parts expliquée et inexpliquée ?

- Part expliquée : c'est la part de la différence totale entre les deux groupes qu'on peut relier à une composition en termes de caractéristiques différente entre les deux groupes.

- Part inexpliquée : l'écart résiduel, idéalement c'est un *effet pur* de l'appartenance au groupe, c'est-à-dire
  - Dans le cas H/F, discrimination de la part des employeurs,
  - Dans le cas public/privé, effet "causal" du secteur.

  mais en pratique l'écart inexpliqué peut être plus complexe à interpréter.

Utilisation d'un contrefactuel

- Pour distinguer expliqué et inexpliqué, on utilise un **contrefactuel** - la valeur que prendrait la variable d'intérêt sous certaines conditions de composition et de valorisation.

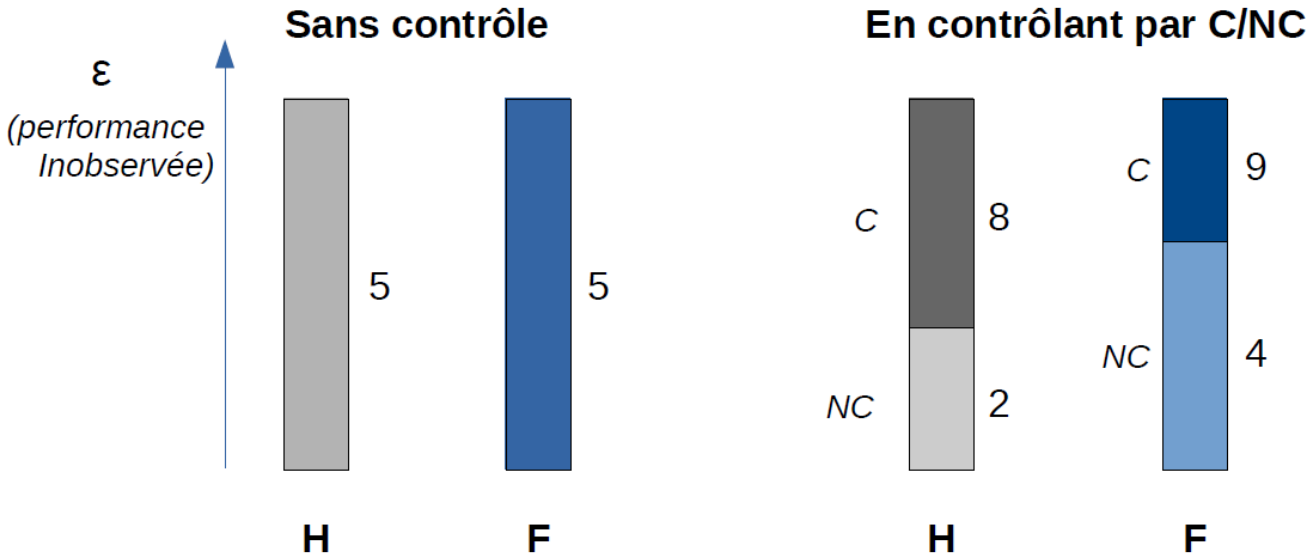
- Une multitudes de façons de choisir et de construire cet objet.

- Celui-ci a un sens sous l'hypothèse essentielle d'**indépendance conditionnelle**. Elle impose que conditionnellement à une composition, les caractéristiques inobservables sont indépendantes du groupe d'appartenance.

- Idée : si les deux groupes ont des caractéristiques inobservables différentes après contrôle des observables, isoler valorisation et composition est impossible.

Interpréter l'écart inexpliqué avec précaution

- Le choix des variables est déterminant : **pour un ensemble donné de caractéristiques observables, les individus des deux groupes sont-ils vraiment comparables ?**
- De façon générale, il y aura toujours une suspicion sur le fait que les caractéristiques inobservées diffèrent (HIC non vérifiée) : parler de *discrimination* ou d'effet causal avec beaucoup de précaution.
- Surtout quand les explicatives reposent sur des choix ou de la sélection (discriminatoires, ou pas). Exemple : statut de cadre.



La décomposition des écarts de moyenne

Principe de la décomposition d'Oaxaca-Blinder

Exemple-type

On étudie l'écart de salaire moyen entre hommes et femmes, en lien avec le fait que ces sous-populations présentent des caractéristiques différentes

On écrit l'équation de salaire pour les hommes d'une part, les femmes d'autre part :

$$Y_i = \beta_{H0} + \sum_{k=1}^{K} X_{ik}\beta_{Hk} + \epsilon_i, \forall i \in H$$

$$Y_i = \beta_{F0} + \sum_{k=1}^{K} X_{ik}\beta_{Fk} + \epsilon_i, \forall i \in F$$

Avec Y_i le (log-)salaire individuel et X_{ik} un vecteur de caractéristiques individuelles, par exemple le diplôme, l'expérience, le temps de travail, le secteur...

Analyser des différences moyennes

Les salaires moyens s'écrivent donc $\overline{Y}_H = \hat{\beta}_{H0} + \sum_{k=1}^{K} \overline{X}_{Hk}\hat{\beta}_{Hk}$ et $\overline{Y}_F = \hat{\beta}_{F0} + \sum_{k=1}^{K} \overline{X}_{Fk}\hat{\beta}_{Fk}$ car on a bien $\sum_{i\in H} \hat{\epsilon}_i = 0$ et $\sum_{i\in F} \hat{\epsilon}_i = 0$.

Décomposition d'Oaxaca-Blinder :

$$\underbrace{\overline{Y_H} - \overline{Y_F}}_{\hat{\Delta}_O} = \hat{\beta}_{H0} + \sum_{k=1}^{K} \overline{X_{Hk}}\hat{\beta}_{Hk} - \hat{\beta}_{F0} - \sum_{k=1}^{K} \overline{X_{Fk}}\hat{\beta}_{Fk}$$

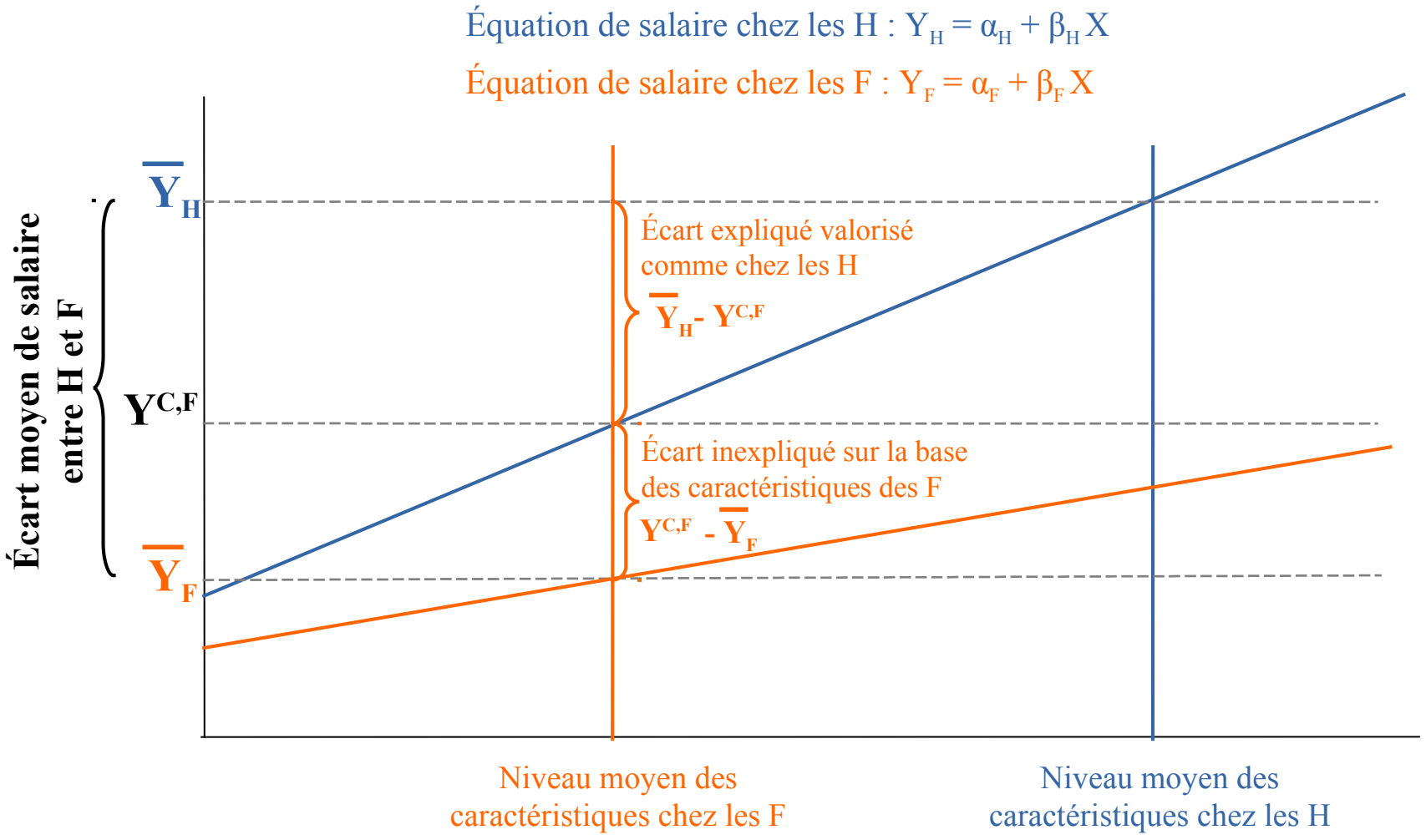
$$= \underbrace{\left(\hat{\beta}_{H0} - \hat{\beta}_{F0}\right) + \sum_{k=1}^{K} \overline{X_{Fk}}\left(\hat{\beta}_{Hk} - \hat{\beta}_{Fk}\right)}_{\hat{\Delta}_S \quad \text{(inexpliqué)}} + \underbrace{\sum_{k=1}^{K} \left(\overline{X_{Hk}} - \overline{X_{Fk}}\right)\hat{\beta}_{Hk}}_{\hat{\Delta}_X \quad \text{(expliqué)}}$$

$\hat{\Delta}_O$: écart de salaire observé ou écart brut

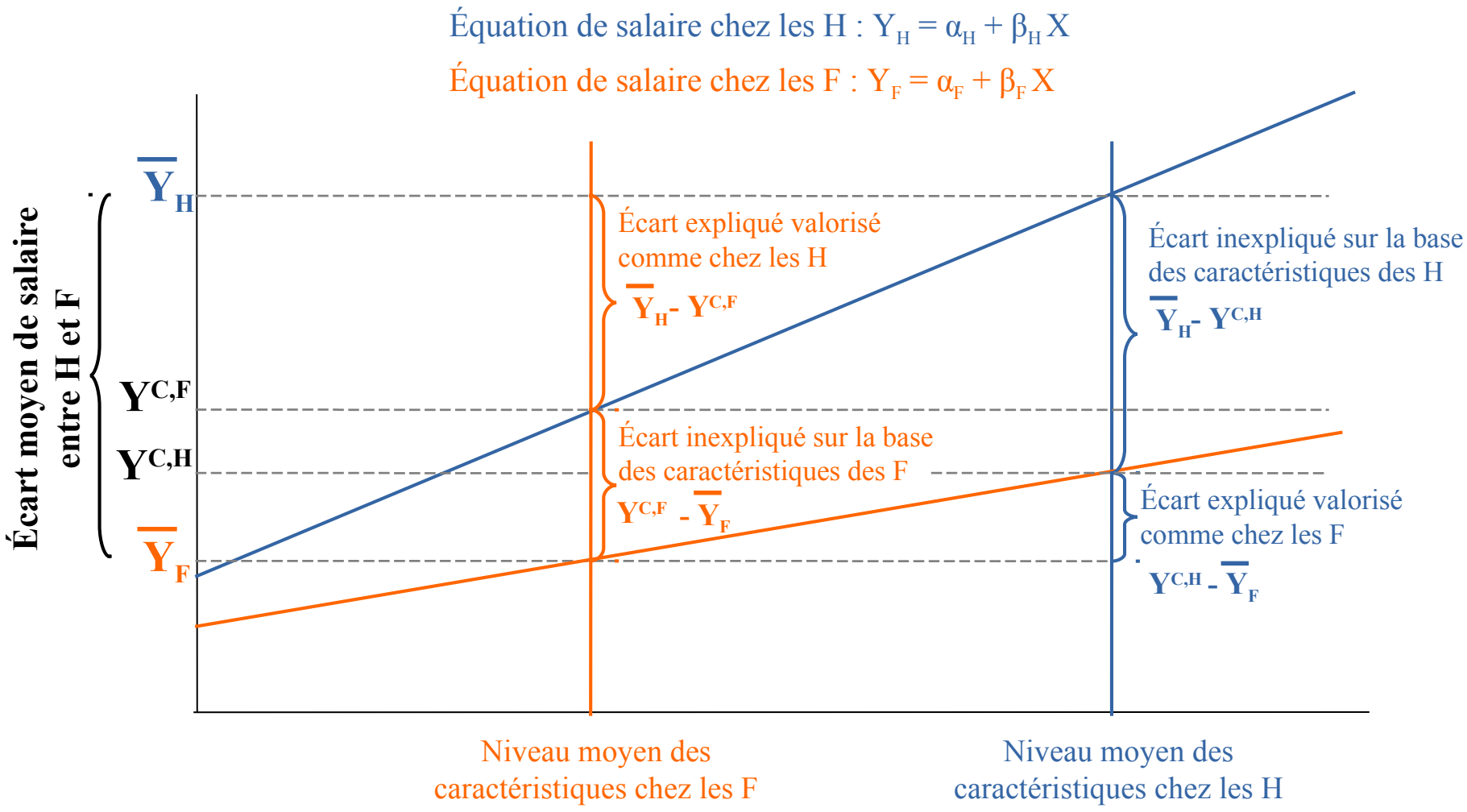
$\hat{\Delta}_X$: composante expliquée ou effet de composition

$\hat{\Delta}_S$: composante inexpliquée ou *wage structure effect*

Illustration avec une constante et une variable continue



Deux contrefactuels simples



Décomposition agrégée ou détaillée

Dans le cas simple d'Oaxaca-Blinder, il est facile de procéder à une décomposition détaillée c'est-à-dire de mesurer plus seulement $\hat{\Delta}_X$ mais les termes de la somme :

$$\hat{\Delta}_X = \sum_{k=1}^{K} \hat{\Delta}_{X_k}$$

où pour chaque covariable X_k, $\hat{\Delta}_{X_k}$ désigne sa contribution à l'écart expliqué : $\hat{\Delta}_{X_k} = \left(\overline{X_{Hk}} - \overline{X_{Fk}}\right) \hat{\beta}_{Hk}$.

Sous des hypothèses supplémentaires, il est également possible d'identifier les contributions des variables à l'écart inexpliqué : $\hat{\Delta}_S = \sum_{k=0}^{K} \hat{\Delta}_{S_k} = \sum_{k=0}^{K} \overline{X_{Fk}} \left(\hat{\beta}_{Hk} - \hat{\beta}_{Fk}\right)$.

La décomposition d'Oaxaca-Blinder

Ce qui est décomposé	La différence de deux moyennes d'une variable continue.
Contrefactuel	Modèle linéaire estimé dans un groupe, appliqué aux X de l'autre, typiquement, le niveau de salaire moyen des femmes si leurs caractéristiques étaient valorisées comme celles des hommes.
Avantages / inconvénients	Simple, disponible dans les principaux logiciels, décomposition détaillée immédiate, mais limite l'analyse à la seule moyenne.

Extension au cas binaire : la décomposition de Fairlie

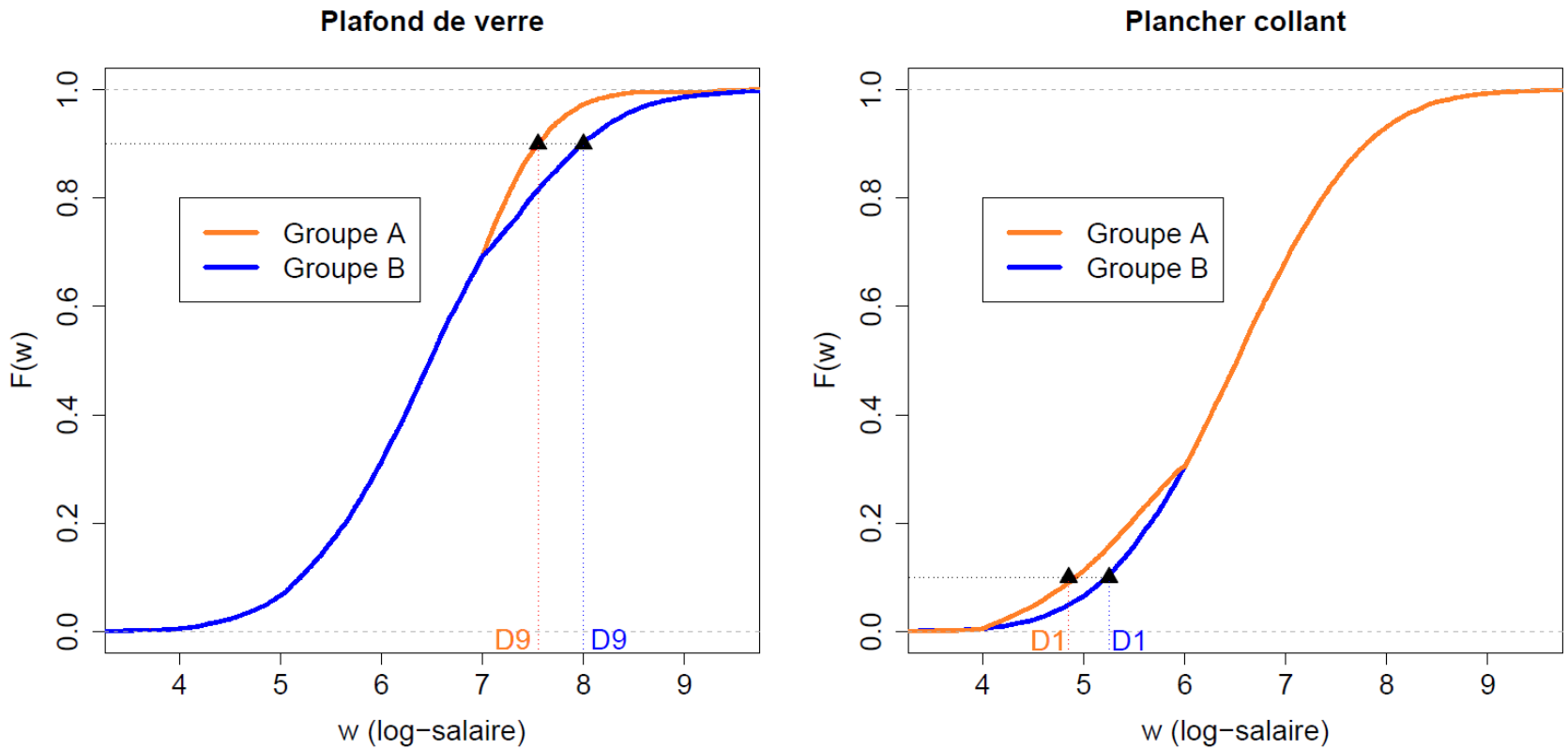
Fairlie (1999), "The Absence of the African-American Owned Business: An Analysis of the Dynamics of Self-Employment".

Ce qui est décomposé	La différence de deux moyennes d'une variable catégorielle : de deux probabilités.
Contrefactuel	Modèle logit ou probit estimé dans un groupe, appliqué aux X de l'autre
Avantages / inconvénients	Simple pour la décomposition agrégée, plus compliquée pour la décomposition détaillée et soumis à des hypothèses paramétriques sur les erreurs.

La décomposition de l'écart entre distributions

Décompositions au-delà de la moyenne

Exemple : permet d'étudier des phénomènes de type "plafond de verre" ou "plancher collant".



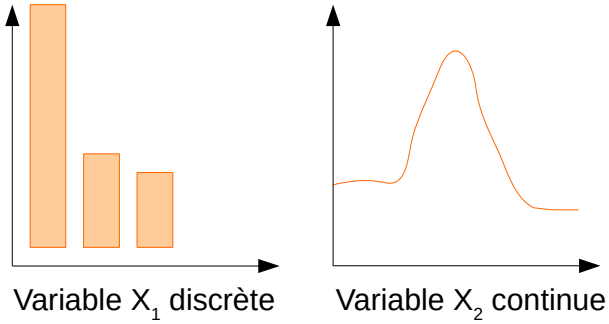
Décompositions au-delà de la moyenne

- On s'intéresse à présent à **n'importe quelle statistique $\nu(F)$ de la distribution d'une variable continue**.
  - Les quantiles : médiane, déciles, quartiles...
  - Toute autre statistique construite à partir de la distribution : rapport inter-déciles, écart-type, coefficient de Gini, fonction d'accès...

- Sachant que Y dépend d'un ensemble de caractéristiques X, la distribution (non-conditionnelle ie. effectivement observée) de Y dans une population H donnée dépend de :
  - La distribution des caractéristiques X dans la population H
  - La distribution de Y sachant X (distribution conditionnelle de Y)

Des X à la distribution des Y

Distribution des X dans le groupe F

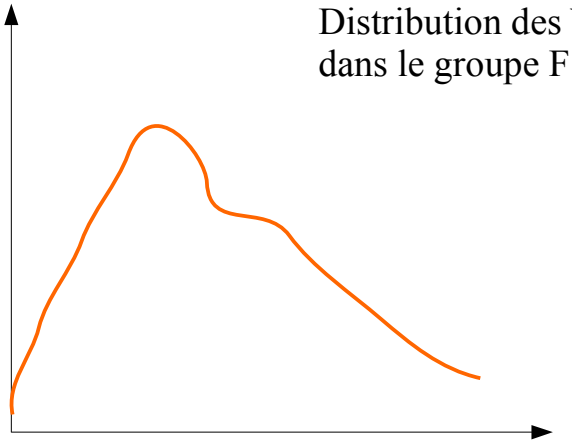


$F_{Y_F \mid X}$

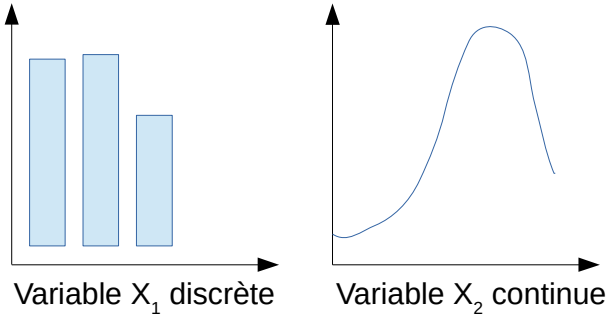
Distribution conditionnelle des Y sachant les X dans le groupe F

$F_{Y_F \mid X_F} = F_{Y_F}$

Distribution des Y dans le groupe F



Distribution des X dans le groupe H

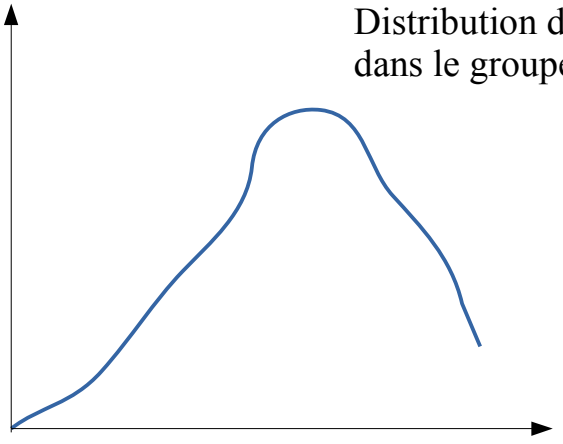


$F_{Y_H \mid X}$

Distribution conditionnelle des Y sachant les X dans le groupe H

$F_{Y_H \mid X_H} = F_{Y_H}$

Distribution des Y dans le groupe H



Comment faire ?

- On cherche à décomposer l'écart entre la distribution $F_{Y_F}(= F_{Y_F|X_F})$ et la distribution $F_{Y_H}(= F_{Y_H|X_H})$.

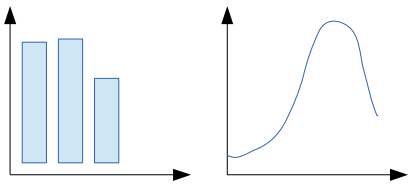
- On utilise comme contrefactuel la distribution de salaire des femmes si leurs caractéristiques étaient valorisées de la même manière que celles des hommes, soit $F_{Y_H|X_F}$ résultant de :
  - La distribution conditionnelle du groupe H, soit $F_{Y_H|X}$ ...
  - ... appliquée aux caractéristiques X du groupe F

(Plus formellement, on a $F_{Y_H|X_F} = \int F_{Y_H|X}(y|X) dF_{X_F}(X)$.)

Deux approches pour parvenir à un contrefactuel

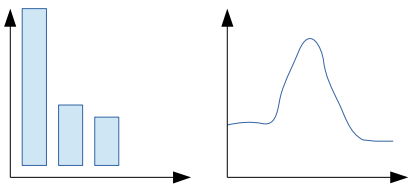
Repondération

Distribution des X dans le groupe H

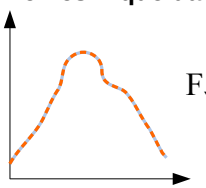


Repondération

Distribution des X dans le groupe H **repondérée pour coïncider avec celle chez les F**

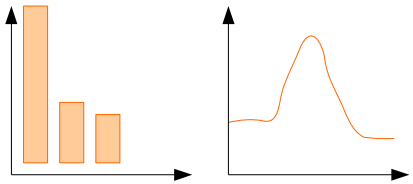


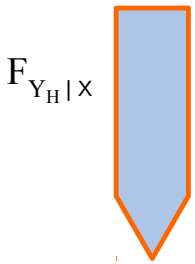
On en déduit la **distribution des Y du groupe H si celui-ci avait les mêmes X que dans le groupe F**



Estimation de la distribution conditionnelle

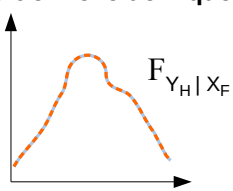
Distribution des X dans le groupe F





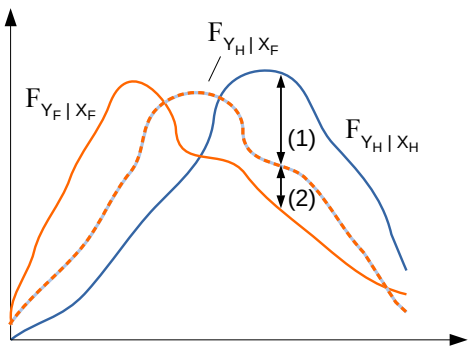
On applique aux X du groupe F la distribution conditionnelle des Y chez les H

Distribution des Y du groupe F si celui-ci avait la même distribution conditionnelle de Y que dans le groupe H



Les deux méthodes estiment la même distribution contrefactuelle (mais peuvent donner des résultats numériques différents).

Celui-ci permet de calculer les parts expliquée (1) et inexpliquée (2) de l'écart en tout point de la distribution.



Méthodes de repondération

Pour arriver à $F_{Y_H|X_F}$, on part de F_{Y_H} et on ajuste la répartition des X (en passant de X_H à X_F), càd on repondère les observations du groupe H afin de rendre leurs caractéristiques X identiques à celles du groupe F.

- Exemple : on veut comparer les distributions de salaires H/F, en contrôlant de l'encadrement. Les femmes exercent moins souvent des fonctions d'encadrement.

- On va repondérer :
  - à la baisse les observations H avec des fonctions d'encadrement,
  - à la hausse les H sans fonctions d'encadrement.

Les méthodes de repondération

DiNardo, Fortin et Lemieux (1996), "Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach", *Econometrica*.

Ce qui est décomposé	La différence de deux statistiques d'une distribution.
Contrefactuel	Distribution contrefactuelle obtenue en pondérant les observations d'un groupe de sorte qu'ils miment la distribution des X de l'autre groupe.
Avantages / inconvénients	Simple pour la décomposition agrégée, plus compliquée pour la décomposition détaillée - non additive ou dépendante de l'ordre, attention au problème de support commun.

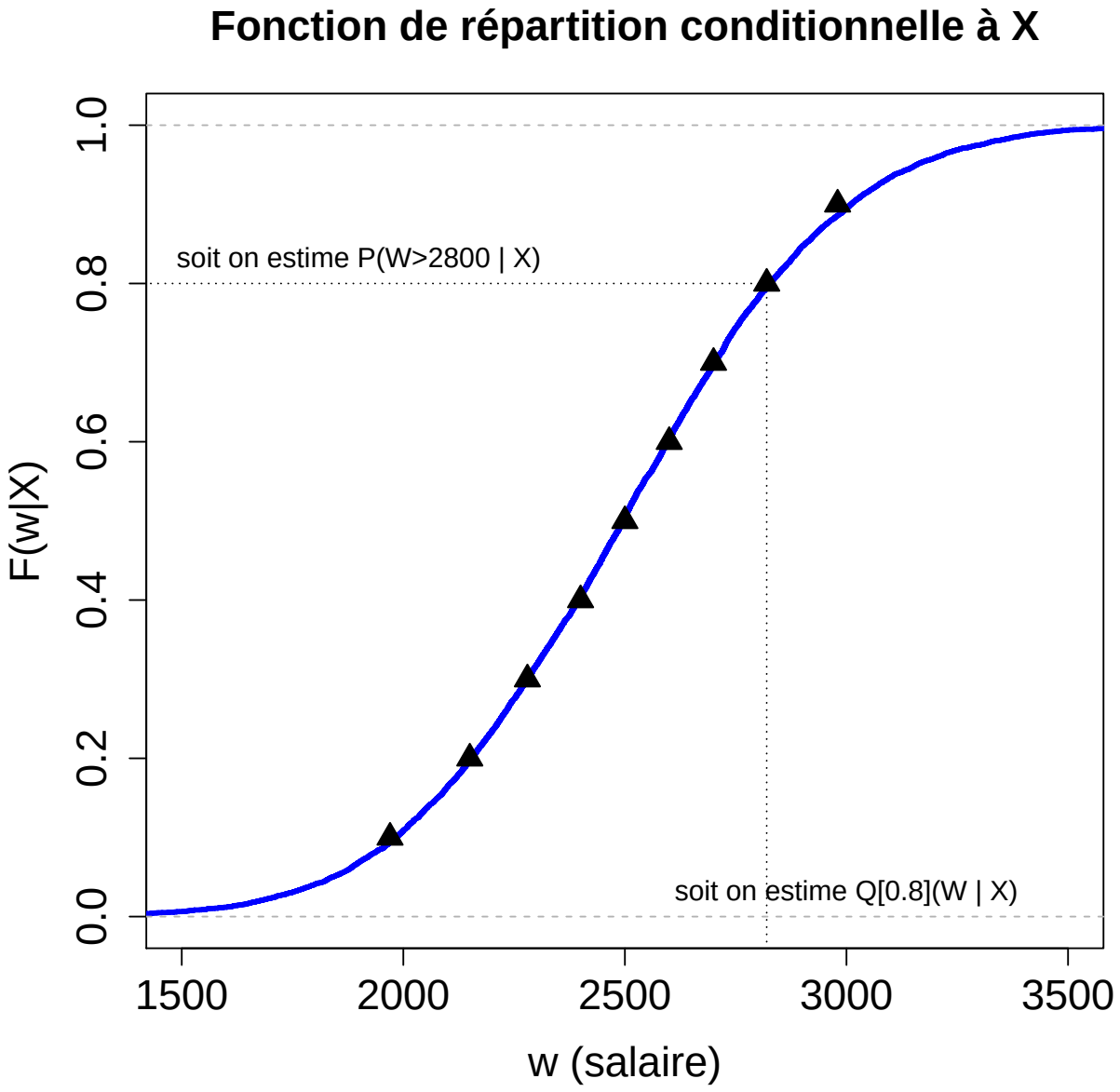
Estimer la distribution conditionnelle

Pour arriver à $F_{Y_H|X_F}$, on part de F_{X_F} et on remplace la distribution conditionnelle $F_{Y_F|X}$ par $F_{Y_H|X}$, ce qui nécessite d'estimation la distribution conditionnelle des Y dans le groupe des H.

Deux façons de réaliser l'estimation :

- "verticalement" : régression sur distribution
  - On estime $P(W > 2800|X)$, $P(W > 2850|X)$, $P(W > 2900|X)$ ... en balayant l'ensemble des salaires.
  - Chernozhukov Fernandez-Val et Melly (2013).
- "horizontalement" : estimation des quantiles conditionnels
  - On estime $Q_{0.8}(W|X)$, $Q_{0.81}(W|X)$... en balayant $[0, 1]$.
  - Machado et Mata (2005), Melly (2006).

Estimer la distribution conditionnelle - deux approches



Estimer la distribution conditionnelle - des calculs intensifs

- Dans tous les cas, on doit estimer un grand nombre de régresssions pour reconstituer l'ensemble de la fonction de distribution, puis appliquer chacun de ces modèles conditionnels (des H) aux caractéristiques X des F.

- Procédure très intensive en calculs : un grand nombre de régressions (logit/probit ou quantile) est nécessaire pour balayer toute la distribution.

Les méthodes d'estimation de la distribution conditionnelle

Chernozhukov, Fernández-Val, et Melly (2013), "Inference on counterfactual distributions", *Econometrica*.

Machado et Mata (2005), "Counterfactual decomposition of changes in wage distributions using quantile regression", *Journal of applied Econometrics*.

Ce qui est décomposé	La différence de deux statistiques d'une distribution.
Contrefactuel	Distribution contrefactuelle obtenue appliquant aux X d'un groupe la distribution conditionnelle estimée dans l'autre groupe.
Avantages / inconvénients	Plus performant quand les X jouent de façon spécifique à certains points de la distribution des Y, très intensif en calcul, la décomposition détaillée est là aussi non additive ou dépendante de l'ordre.

Les régressions RIF

RIF - Une généralisation d'Oaxaca-Blinder au cas quantile

- L'idée est de se ramener au cadre linéaire simple d'Oaxaca-Blinder, qui permet notamment d'obtenir des décompositions détaillées additives et invariantes de l'ordre.

- Pour ce faire, on utilise un type de régression quantile tel que :

$$\hat{Q}_\tau = \bar{X}\hat{\gamma}_\tau.$$

- Attention !! Ce ne sont pas les régressions quantiles "usuelles" à la Koenker et Bassett (1978), où l'on modélise le quantile de Y conditionnel à X. Ici on s'intéresse au quantile non-conditionnel.

RIF - Les régressions RIF

- On utilise la *fonction d'influence recentrée*. Celle-ci mesure la façon dont une observation de Y influence une statistique donnée (en l'occurrence, le quantile non conditionnel).
- Dans le cas quantile, pour un seuil τ la RIF vaut :

$$RIF(Y_i; Q_\tau) = Q_\tau + \frac{\tau - \mathbf{1}\{Y_i \leqslant Q_\tau\}}{f_Y(Q_\tau)}$$

- **Propriété essentielle :** $\mathbb{E}(RIF(Y; Q_\tau) \mid X) = Q_\tau$.
- Une régression RIF correspond tout simplement à une régression par MCO de $RIF(y; Q_\tau)$ sur X.

Les régressions RIF

En pratique, pour une décomposition au $\tau^{\text{ème}}$ quantile :

(1) On transforme chaque Y_i en $\hat{RIF}(Y_i, Q_{g,\tau})$

(2) On régresse les $\hat{RIF}(Y_i, Q_{g,\tau})$ sur les X_i par MCO :

$$\hat{RIF}(Y, Q_{H,\tau}) = X\gamma_H + \epsilon_H$$

$$\hat{RIF}(Y, Q_{F,\tau}) = X\gamma_F + \epsilon_F$$

(3) Les $\hat{\gamma}_{H,\tau}$ et $\hat{\gamma}_{F,\tau}$ obtenus sont utilisés pour une décomposition de type Oaxaca-Blinder :

$$\underbrace{\mathbb{E}(RIF(Y; Q_{H,\tau})) - \mathbb{E}(RIF(Y; Q_{F,\tau}))}_{Q_{H,\tau} - Q_{F,\tau} = \hat{\Delta}_O^\tau} = \underbrace{(\hat{\gamma}_{H,\tau} - \hat{\gamma}_{F,\tau})\bar{X}_F}_{\hat{\Delta}_S^\tau} + \underbrace{(\bar{X}_H - \bar{X}_F)\hat{\gamma}_{H,\tau}}_{\hat{\Delta}_S^\tau}$$

Les régressions RIF

Firpo, Fortin et Lemieux (2009), "Unconditional quantile regressions", *Econometrica*.

Ce qui est décomposé	La différence des fonctions d'influence centrées d'une statistique, ce qui revient à la différence des statistiques entre les groupes.
Contrefactuel	Fonction d'influence centrée si les caractéristiques d'un groupe étaient valorisées comme celles de l'autre groupe.
Avantages / inconvénients	On se ramène au cadre Oaxaca-Blinder : simplicité, facilité de la décomposition détaillée. Dépend de la qualité de l'approximation linéaire - problématique en cas de points de masse.

Conclusion

- Une grande variété d'applications possibles (cf. annexes pour des exemples sur données françaises).
- Méthodes de repondération qui permettent d'étendre facilement l'approche des méthodes de décomposition à d'autres statistiques d'intérêt que la seule moyenne.
- Mais des précautions d'usage pour interpréter les résultats : l'hypothèse d'indépendance conditionnelle est essentielle pour assimiler l'inexpliqué à une discrimination. Or, celle-ci est probablement invalide dès que les explicatives relèvent d'un mécanismes de (auto-)sélection.

Merci de votre attention !

Exemples d'application - Oaxaca-Blinder

- E. Coudin, S. Maillard et M. Tô (2017), "Ecarts salariaux entre les entreprises et au sein de l'entreprise : femmes et hommes payés à la même enseigne ?", *Insee Références.*

- D. Audenaert, J. Bardaji, R. Lardeux, M. Orand et M. Sicsic (2014), "La résistance des salaires depuis la grande récession s'explique-t-elle par des rigidités à la baisse ?", *L'économie française.*

- M. Meunier (2007), "Origine migratoire et performance scolaire : décomposition des scores PISA 2000", *Université de Genève*.

Exemples d'application - Méthodes de repondération et RIF

- C. Bonnet, A. Keogh et B. Rapoport (2014), "Quels facteurs pour expliquer les écarts de patrimoine entre hommes et femmes en France ?", *Economie et statistique*.
- E. Coudin et A. Souletie (2016), "Obésité et marché du travail : les impacts de la corpulence sur l'emploi et le salaire", *Economie et Statistique*.
- B. Boutchenik et J. Lê (2017), "Les descendants d'immigrés maghrébins : des difficultés d'accès à l'emploi et aux salaires les plus élevés", *Insee Références*.

- C. Champagne, A. Pailhé et A. Solaz (2015), "Le temps domestique et parental des hommes et des femmes : quels facteurs d'évolutions en 25 ans ?", *Economie et Statistique*.
- A. Boiron (2016), "Evolution des inégalités de niveau de vie entre 1970 et 2013", *Insee références*.

Repondération - quels poids utiliser ?

On repondère une observation H dont les caractéristiques sont X_i par :

$$\Psi(X_i) = \frac{P(X_i|F)}{P(X_i|H)},$$

On peut voir que $\Psi(X) = \frac{P(X|F)}{P(X|H)} = \frac{P(F=1|X)}{P(H=1|X)}[\frac{P(H)}{P(F)}]$.

(1) On estime la probabilité d'appartenance au groupe F sachant X $P(F = 1|X)$ dans tout l'échantillon.

(2) On calcule pour chaque observation $i \in H$, $\hat{\Psi}(X_i)$ en remplaçant les grandeurs empiriques correspondantes :
- $\hat{P}(F = 1|X_i)$: probabilité prédite obtenue à l'étape (1)
- $\hat{P}(H = 1|X_i) = 1 - \hat{P}(F = 1|X_i)$
- $\hat{P}(F)$ et $\hat{P}(H)$: proportions dans l'échantillon

(3) On calcule la statistique d'intérêt $\nu()$ sur les observations H repondérées par $\hat{\Psi}(X_i)$: celle-ci correspond à $\hat{\nu}(F_{Y_H|X_F})$.